Phyla_AMPHORA User Manual

A Phylum-specific Automated Phylogenomic Inference Pipeline for Bacterial Sequences.

COPYRIGHT
2012 by Martin Wu

Phyla_AMPHORA is free software: you may redistribute it and/or modify its under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or any later version.

Phyla_AMPHORA is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details (http://www.gnu.org/licenses/).

For any other inquiries send an Email to Martin Wu: mw4yv@virginia.edu

CITATION
When publishing work that is based on the results from Phyla_AMPHORA, please cite:

Wang Z and Wu M: A Phylum-level Bacterial Phylogenetic Marker Database. Mol. Biol. Evol. Advance Access publication March 21, 2013.
doi:10.1093/molbev/mst059

DEPENDENCY

Phyla_AMPHORA depends on several external programs.

1. HMMER3 (http://hmmer.janelia.org/). Required for marker identification, sequence alignment and trimming. Earlier versions of HMMER will not work!
2. RAxML version 7.3.0 or later (https://github.com/stamatak/standard-RAxML/downloads). Required for phylotyping.
3. Bioperl 1.5.2 or later (http://www.bioperl.org/wiki/Getting_BioPerl).
4. EMBOSS (http://emboss.sourceforge.net/download/). The 'getorf' program of the EMBOSS package is required only if you analyze DNA sequences using Phyla_AMPHORA

Make sure that these programs are installed and are in your system's executable search path. To test, in a terminal type

raxmlHPC -version
raxmlHPC-PTHREADS -version
hmmsearch -h
hmmalign -h
getorf -help

If you see version or help messages, then these programs have been correctly installed. It is important to make sure they are the correct versions.

A script named 'preinstall.pl' is also included with Phyla_AMPHORA to check and install the dependencies automatically. You need the privilege of the system administrator to run the script. See below for instructions.

INSTALLATION

1. Download Phyla_AMPHORA
2. Unpack Phyla_AMPHORA
   tar -zxvf Phyla_AMPHORA.tar.gz
3. Install dependencies if they have not been installed
   cd Phyla_AMPHORA
   sudo perl preinstall.pl
4. Setup Phyla_AMPHORA.
   You need to set up the environment variable 'Phyla_AMPHORA_home' so the Phyla_AMPHORA scripts know where to look for the phylogenetic marker database and the NCBI taxonomy information. Let's suppose your unpacked Phyla_AMPHORA folder is at /home/foo/Phyla_AMPHORA.

   If you are using a bash shell, you can add the following lines to the end of the file ~/.bashrc

export Phyla_AMPHORA_home=/home/foo/Phyla_AMPHORA

Then in the terminal, issue this command

source ~/.bashrc

If you are using a C shell, you can add the following lines to the end of the file ~/.tcshrc.

setenv Phyla_AMPHORA_home /home/foo/Phyla_AMPHORA

Then in the terminal, issue this command

source ~/.tcshrc

5. Make the Phyla_AMPHORA scripts executable.

*chmod +x /home/foo/Phyla_AMPHORA/Scripts/**

You should see five folders.

1. Marker

This folder contains a seed alignment file in Stockholm format (*.stock), an alignment mask file (*.mask), a profile HMM file (*HMM) and a tree file in newick format (*.tre) for each marker gene.

For more information about the phylogenetic markers that are included in Phyla_AMPHORA, see the marker.list file in the Marker folder.

IMPORTANT: Because the Marker folder exceeds the 1GB (the size limit of github), it is not included in the github package. If you download Phyla_AMPHORA from github, you should download the Marker database from http://wolbachia.biology.virginia.edu/WuLab/Software.html and move it here.

2. Scripts

This folder contains the scripts for marker identification, alignment, trimming and phylotyping.

3. Taxonomy

This folder contains the NCBI taxonomy database that is used by the Phylotyping.pl script for phylotyping.

4. Tree

This folder contains the genome trees for 20 bacterial phyla in newick format. The genome trees are RAxML maximum likelihood trees made from concatenated protein sequences of the phylum-specific markers.

5. TestData

This folder contains the E. coli genome assembly (ecoli.fasta) and proteome sequences (ecoli.pep) for testing Phyla_AMPHORA.

RUNNING Phyla_AMPHORA

We recommend that you allocate at least 4GB of memory to Phyla_AMPOHRA.

1. Marker identification

Use MarkerScanner.pl to identify phylum-specific bacterial marker sequences. Given a sequence file, this program will identify markers from the input sequences and generate a protein fasta file for each marker gene in your working directory. For example, Acidobacteria.102.pep, Aquificae.33.pep. When DNA sequences are used as input, this program first identifies ORFs longer than 100 bp in all six reading frames, then scans the translated peptide sequences for the phylogenetic markers.

Usage: perl MarkerScanner.pl <options> sequence-file

Options:

-Phylum:0. All (Default)
1. Alphaproteobacteria
2. Betaproteobacteria
3. Gammaproteobacteria
4. Deltaproteobacteria
5. Epsilonproteobacteria
6. Acidobacteria
7. Actinobacteria
8. Aquificae
9. Bacteroidetes
10. Chlamydiae/Verrucomicrobia
11. Chlorobi
12. Chloroflexi
13. Cyanobacteria
14. Deinococcus/Thermus
15. Firmicutes
16. Fusobacteria
17. Planctomycetes
18. Spirochaetes
19. Tenericutes
20. Thermotogae

-DNA: input sequences are DNA. Default: no
-Evalue: HMMER evalue cutoff. Default: 1e-7
-ReferenceDirectory: the file directory that contain the reference alignments, hmms and masks.
-Help: print help;

Examples:

1a. Identify phylogenetic markers from the E. coli proteome

perl MarkerScanner.pl -Phylum 3 TestData/ecoli.pep

1b. Identify phylogenetic markers from the E. coli genome assembly

perl MarkerScanner.pl -Phylum 3 -DNA TestData/ecoli.fasta

If Phyla_AMPHORA has been installed correctly, at the end of the run in example 1a or 1b, you should see 294 marker protein sequences (*.pep) in your working directory.

1c. If you want to identify phylogenetic markers of all the 20 phyla from a set of metagenomic sequence reads (e.g., 454 reads).

perl MarkerScanner.pl -DNA -Phylum 0 metagenomic.fasta

2. Marker sequence alignment and trimming

This program will align, mask and trim the marker protein sequences. Output will be aligned/trimmed sequences. For example, Acidobacteria.102.aln, Aquificae.33.aln and their corresponding alignment masks. The alignment masks can be used to weigh the alignment columns with the RAxML's -a option (for untrimmed alignment only).

Usage: perl MarkerAlignTrim.pl <options>

Options:

-Trim: trim the alignment using masks embedded with the marker database. Default: no
-Cutoff: the Zorro masking confidence cutoff value (0 - 1.0; default: 0.4);
-ReferenceDirectory: the file directory that contain the reference alignments, hmms and masks.
-Directory: the file directory where sequences to be aligned are located. Default: current directory
-OutputFormat: output alignment format. Default: phylip. Other supported formats include: fasta, stockholm, selex, clustal
-WithReference: keep the reference sequences in the alignment. Default: no
-Help: print help

Example:

perl MarkerAlignTrim.pl -WithReference -OutputFormat phylip

If Phyla_AMPHORA has been installed correctly, at the end of the run, you should see an alignment file (*.aln) and a mask file (*.mask) for each of the marker gene in your working directory.

It is important to know that in order to run the Phylotyping.pl script properly, the MarkerAlignTrimp.pl needs to be run using '-WithReference -OutputFormat phylip' options.

3. Phylotyping

Use Phylotyping.pl to assign phylotypes for each identified marker sequences. This program will assign each identified marker sequence a phylotype using the parsimony method or the evolutionary placement algorithm of RAxML. The marker sequences need to be aligned first with the reference sequences using MarkerAlignTrim.pl (see above). The alignments should be in the phylip format.

Usage: perl Phylotyping.pl <options>

Options:

-Method: use 'maximum likelihood' (ml) or 'maximum parsimony' (mp) for phylotyping. Default: ml

-CPUs: turn on the multiple thread option and specify the number of CPUs/cores to use. Important: Make sure raxmlHPC-PTHREADs is installed. If the number specified here is larger than the number of cores that are free and available, it will actually slow down the script.

-Help: print help;

If your computer has multiple CPUs/cores, the phylotpying process can be sped up by running multiple threads of the RAxML. However, it is very important to check how many CPUs/cores are free and available to Phylotyping.pl. If you specify a number that is larger than the number of CPUs/cores that are actually available, it will slow down the script. For example, your computer has 8 CPUs but 2 of them are used by other processes. In this case, you can run Phylotyping.pl on 6 CPUs by using the '-CPUs 6' option. Of course, raxmlHPC-PTHREADS needs to be installed.

Example:

Assign phylotypes using the maximum likelihood method

perl Phylotyping.pl -CPUs 6 > phylotype.result

Again, if Phyla_AMPHORA has been installed correctly, you should see something like this as the output:

```
Query   Marker Superkingdom   Phylum  Class   Order   Family  Genus   Species
NP_414730-NC_000913     Gamma.134       Bacteria(1.00)  Proteobacteria(1.00)
Gammaproteobacteria(1.00)       Enterobacteriales(1.00) Enterobacteriaceae(1.00)
Escherichia(1.00)       Escherichia coli(1.00)
NP_417099-NC_000913     Gamma.167       Bacteria(0.96)  Proteobacteria(0.96)
Gammaproteobacteria(0.96)       Enterobacteriales(0.96) Enterobacteriaceae(0.96)
Escherichia(0.70)       Escherichia coli(0.70)
NP_416616-NC_000913     Gamma.252       Bacteria(0.96)  Proteobacteria(0.96)
Gammaproteobacteria(0.96)       Enterobacteriales(0.96) Enterobacteriaceae(0.96)
Escherichia(0.74)       Escherichia coli(0.74)
NP_418155-NC_000913     Gamma.286       Bacteria(0.96)  Proteobacteria(0.96)
Gammaproteobacteria(0.96)       Enterobacteriales(0.96) Enterobacteriaceae(0.96)
Escherichia(0.84)       Escherichia coli(0.84)
```

NP_417422-NC_000913 Gamma.296 Bacteria(0.96) Proteobacteria(0.96) Gammaproteobacteria(0.96) Enterobacteriales(0.96) Enterobacteriaceae(0.96) Escherichia(0.76) Escherichia coli(0.76)
NP_417226-NC_000913 Gamma.306 Bacteria(0.97) Proteobacteria(0.97) Gammaproteobacteria(0.97) Enterobacteriales(0.97) Enterobacteriaceae(0.97) Escherichia(0.61) Escherichia coli(0.61)
NP_417800-NC_000913 Gamma.44 Bacteria(0.95) Proteobacteria(0.95) Gammaproteobacteria(0.95) Enterobacteriales(0.95) Enterobacteriaceae(0.95) Escherichia(0.95) Escherichia coli(0.95)

The phylotyping results are tab-delimited. The numbers within the parentheses are the confidence scores of the assignment.

If you see the following error message when you run Phylotyping.pl, you can delete the 'name2id' file in the folder AMPHORA2_home/Taxonomy/ and run the script again.

---------- EXCEPTION: Bio::Root::Exception ----------
MSG: No such file or directory AMPHORA2_home/Taxonomy/names2id
STACK: Error::throw
STACK: Bio::Root::Root::throw /lib/site_perl/5.16.3/Bio/Root/Root.pm:368
STACK: Bio::DB::Taxonomy::flatfile::_db_connect
/lib/site_perl/5.16.3/Bio/DB/Taxonomy/flatfile.pm:463
STACK: Bio::DB::Taxonomy::flatfile::new /lib/site_perl/5.16.3/Bio/DB/Taxonomy/flatfile.pm:144
STACK: Bio::DB::Taxonomy::new /lib/site_perl/5.16.3/Bio/DB/Taxonomy.pm:116
STACK: AMPHORA2_home/Scripts/Phylotyping.pl:58